**Methods**

**Metagenomic Binning**

The requisite metadata information recommended by the Genomic Standards Consortium for metagenome-assembled genome (1) is shown in `Mystic\_MAG\_Quality\_Stats.xlsx`. The following approaches apply to all bins derived. Taxonomic classification was performed using the *taxator-tk* algorithm (4). Reassembly of bins and initial co-assembly was performed using *SPAdes* 3.11 (5). The initial assembly was completed in metagenomic mode and reassembly was done for each bin in `careful` mode using the first pass contigs as `untrusted contigs`.

Contigs were binned according to their coverage & tetramer frequencies. A set of consensus bins were derived from the bins produced by the *maxbin2*, *metabat2*, and *concoct* algorithms. Completeness and contamination assessment were performed using the lineage workflow in *CheckM*. The bioinformatics pipeline from QC, to assembly, to binning, to refinement, to reassembly, and taxonomic classification was done within the *metaWRAP* software (11). *Prokka* (6) was used to facilitate gene calling and preliminary annotations. Within prokka, Prodigal, *barrnap*, and *Aragorn v1.2* were used to call ORFs, rRNA & tRNA respectively (7,8). All protein sequences were generated using Translation Table 11.

**Metabolic Model Processes**

Additional protein annotations at the nucleotide level were conducted using `metabolic-hmms` collection provided by the Banfield lab (9) and the ` dbCAN-fam-HMMs.v6` collection provided by BioEnergy Science Center of the DOE (10). These sets of HMMs were manually added to the HMM database directory used by Prokka for annotation. Annotations of protein sequences were conducted using KEGG BlastKOALA and GhostKOALA annotation web service (2) and any classifications with a GHOSTX score below 100 were removed, as recommended by Raethong et al. (3).

**Binning Fidelity**

To investigate the consistency of the output of the binning pipeline, as compared to contemporary alternatives (14), the published raw reads collected from a study in the Gulf of Mexico dead zone were processed in parallel with our samples. The bins produced by our pipeline were compared to metagenomic-assembled genomes uploaded and published to the IMG database. Based on initial inspections, the bins produced by *maxbin2* most closely matched those uploaded by Thrash et al.. Furthermore, the refined & reassembled bins produced by *metaWRAP* were restricted to >70% completion and <10% contamination, so the final output of the pipeline did not contain matches to the two smallest genome fragments presented in Thrash et al. It is notable that even those bins that were clearly manually refined i.e. divided into multiple smaller bins and presented as distinct organisms (43-1/2 & 45-1/2) were matched to unique bins produced by *maxbin2*.

*mash* was used to determine the closest matches between bin sets (15). Shared hashes are the units of homology produced by *mash*. Despite having greater sensitivity to mismatches and less accuracy than alignment-based comparisons, it was found to suitable for rapid similarity assessment. The degree of “concentration” or “specificity” of a match was derived from these data and was calculated as the ratio of shared hashes between the query genome and the single best reference genome, divided by the sum of hashes shared with all members of the reference set. This value was calculated for members of a match to determine the degree of dispersion of a given MAG across the entire other bin set. These data, as well as the others described in this section are contained with the `reciprocal\_alignment.tsv` file in the `Data/Thrash\_Libs` directory of the repository.

*mummer* was used to perform genome-to-genome alignments (16). The “--mum” tag was used to ensure only unique alignments on both reference and query were produced. Alignments were filtered to only include regions with longer than 1 Kbp of homology. The genome size and the fraction of the genome aligned for both members of a match are shown in the file mentioned above.

**Linking OTUs from Shotgun Libraries to Bins & Amplicon Profiles**

Both OTUs derived from shotgun libraries, as well as those observed in amplicon libraries prepared for samples collected on 2013-08-12 were assigned to bins based on similarities in abundance and taxonomic classification. Replicate samples were combined by their mean. The minimum observed Pearson correlation between replicates was 0.79. Amplicon OTUs were filtered by the sum of their abundances in comparable samples to the top 500/1000. A total of 11 samples were compared between library types including odd depths from 3 to 17 and 20, 21, and 22 meters. Abundances were normalized using the L1 norm along both axes, within each composition vector first and then within each amplicon/bin.

Raw abundance (*A*) of a bin in a given sample *j* were calculated within metaWRAP using the following equation:

The product of the length of contig (*li*) and its coverage (*cij*) in sample *j* was summed across all *N* contigs assigned to the bin and then divided by total length of the bin genome. These raw values were divided by the number of reads in each library before following the same normalization steps applied to OTU abundances.

OTU taxonomy was assigned using the RDP classifier using the 16s rRNA Training Set 16 (12). Confidence scores below 50% and *Incertae Sedis* classifications were removed from the “fixed rank” output. Each level in the hierarchy between Kingdom & Genus was treated as an independent feature to match the format of the taxonomic assignments produced by metaWRAP.

Pairwise Euclidean distances between abundance vectors for bins and OTUs were calculated and seen to range from 0 to 1 on L1 normalized vectors. The fraction of matching taxonomic labels was then subtracted from the distance to produce a combined distance metric ranging from 1 to -1, where the latter represents a perfect match. The tolerance of this metric for producing correct matches was a challenge to assess with the dearth of known matches usable for scoring assignments. Eight bins were used for this purpose as these 5 of them received an assembled 16S rRNA gene copy and the other three were positive control organisms. The tolerance was trained on these bins and was observed to be different for the amplicon library data (0.0251), as compared to shotgun library data (0.176). The preprocess and assignment routines are outlined in the `assign\_bin\_16s.py` script and in annotated form in the `16S\_to\_Bin\_Matching.ipynb` Jupyter notebook.

**Gene Abundance Quantification**

The abundances for select genes was annotated with biogeochemical process-related KOs. Salmon was used to quantify the abundances of these sequences in each library (13). The strict nucleotide sequence of each gene was bookended by 50-100 bp on each end to ensure that near identical gene sequences on distinct contigs were not conflated. Uniqueness was confirmed during indexing, as Salmon detects and removes duplicates prior to mapping. The entire workflow used to integrate the various sequence and gene feature files produced by Prokka, along with the functional & taxonomic annotation files produced by GHOSTKoala to eventually produce the abundance matrix is outlined in the `salmon\_quant.sh` script.

1. Bowers, R. M. et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nature Biotechnology 35(725–731) doi:10.1038/nbt.3893
2. Kanehisa, M., Sato, Y., and Morishima, K. (2016) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J. Mol. Biol. 428, 726-731 (<http://www.kegg.jp/ghostkoala/>)
3. Raethong, N., et. al. (2016) Sequence- and Structure-Based Functional Annotation and Assessment of Metabolic Transporters in Aspergillus oryzae: A Representative Case Study. BioMed Research International <https://www.hindawi.com/journals/bmri/2016/8124636/>
4. J. Dröge, I. Gregor, and A. C. McHardy. Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. Bioinformatics 2015 31: 817-824.doi: 10.1093/bioinformatics/btu745
5. Bankevich A., et al. (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology. 19.5 (2012): 455-477.
6. Seemann T. (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics. 15;30(14):2068-9.
7. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119. doi:10.1186/1471-2105-11-119.
8. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Research. 2004;32(1):11-16. doi:10.1093/nar/gkh152.
9. Anantharaman, K. et al. (2016) Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nature Communications. 13219. doi:10.1038/ncomms13219
10. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Research. 2012;40(Web Server issue):W445-W451. doi:10.1093/nar/gks479.
11. Uritskiy, G.V. et al. (2018) MetaWRAP - a flexible pipeline for genome-resolved metagenomic data analysis. BioRxiv doi: <https://doi.org/10.1101/277442>
12. Cole, J. R., Chai, Tiedje, J. M., et al. (2006). The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic acids research, 35(1), D169-D172.
13. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature methods, 14(4), 417.
14. Thrash, JC. (2017) Metabolic Roles of Uncultivated Bacterioplankton Lineages in the Northern Gulf of Mexico “Dead Zone”. mBio. 8(5) e01017-17
15. Ondov, B.P et. al (2016) Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology. 2016 17:132 <https://doi.org/10.1186/s13059-016-0997-x>
16. Marçais, G, et al. "MUMmer4: A fast and versatile genome alignment system." PLoS Computational Biology 14.1 (2018): e1005944.